Sampling distributions beyond the z

Statistics for Social Scientists I

Bur, GJM

2024-07-28

1 Overview: sampling distributions generally

So far, we've mastered a couple of specific sampling distributions, those for sample totals and means. Totals are binomially distributed, while means are Normally-distributed. However, there are two limitations on these approaches.

First, we have been assuming that we know the population standard deviation σ under the null hypothesis. That is only true in the case of proportions since the mean of an individual dummy is p and the standard deviation of a individual dummy variable is just $\sqrt{p \cdot (1-p)}$, i.e. it depends on the mean only. Otherwise, it's usually not true, but fixing it does not affect our procedures much.

Second, we've been assuming that all data we have come from simple random sampling processes with no measurement error, no response bias, no clustering, no stratification, no (unintended) wording effects, and so on. This is not true, and it's harder to fix. *In the interest of time, I do not discuss these problems here. You can find a fairly comprehensive discussion, not required for this course, in my notes here.

One theme throughout these notes is that conventional sources really drag out the differences between a bunch of very closely-related sampling distributions; in my view, the six chapters $18-23$ in Moore, Notz, and Fligner should be at most two different chapters. *You should really try to approach this material as bunch of small variations on a theme*, which our book unfortunately does nothing to really make clear.

2 Inference on a single mean which is t-distributed

Let's start by discussing what happens when our population standard deviation is unknown. Then, we use σ to estimate it. We now have a t-statistic:

$$t = \frac{\bar{y} - \mu_0}{\frac{s}{\sqrt{n}}}$$

This t-statistic has a t-distribution, which is approximately Normal, and the difference is only noticeable with small sample sizes. *This results from the fact that we now have a* ***ratio***

of random variables (note the $\bar{y}$ in the numerator and the s in the denominator), which is much more complicated ***in theory*** *than just a single random variable.* In *practice*, the difference is very minimal. How do we adjust for this? Well, we look to the sample's ***degrees of freedom***, a term we've encountered before.

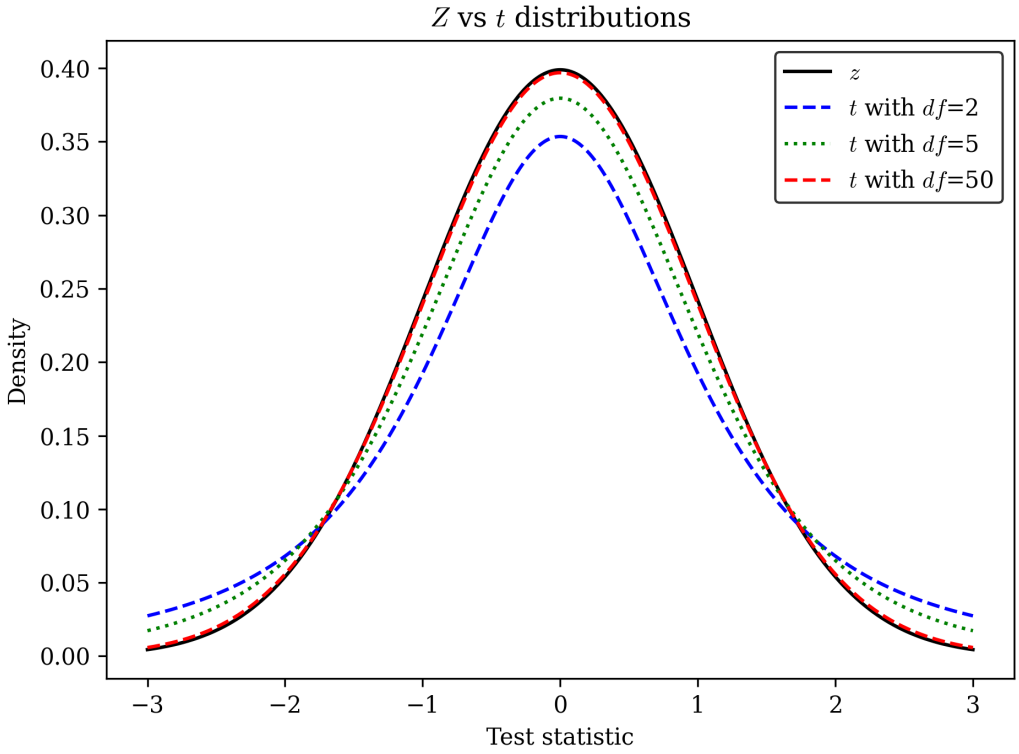
2.1 Degrees of freedom, quickly

"Degrees of freedom" refers to the number of pieces of information that are allowed to vary in the solution of a problem. For example, if I tell you that the mean age in our class is 21, and then I give you the ages of $n-1$ people, you know the age of the nth person immediately. Similiarly, if I do not know the population variance and will use the sample mean to estimate the sampling variance, this means that I'm sacrificing, as it were, one observation since I am not calculating the variance from n observations that are totally free to vary but rather from $n-1$ since the observations must sum to the sample mean (this is not true if I use the population mean in calculating the sampling variance). In general, in estimating some parameter θ, $df = \nu = n - n_k$ where we are estimating k other parameters in order to get an equation for θ. Note that the symbol ν is Greek "new" and makes the n sound. I won't use it often, but you might see it elsewhere.

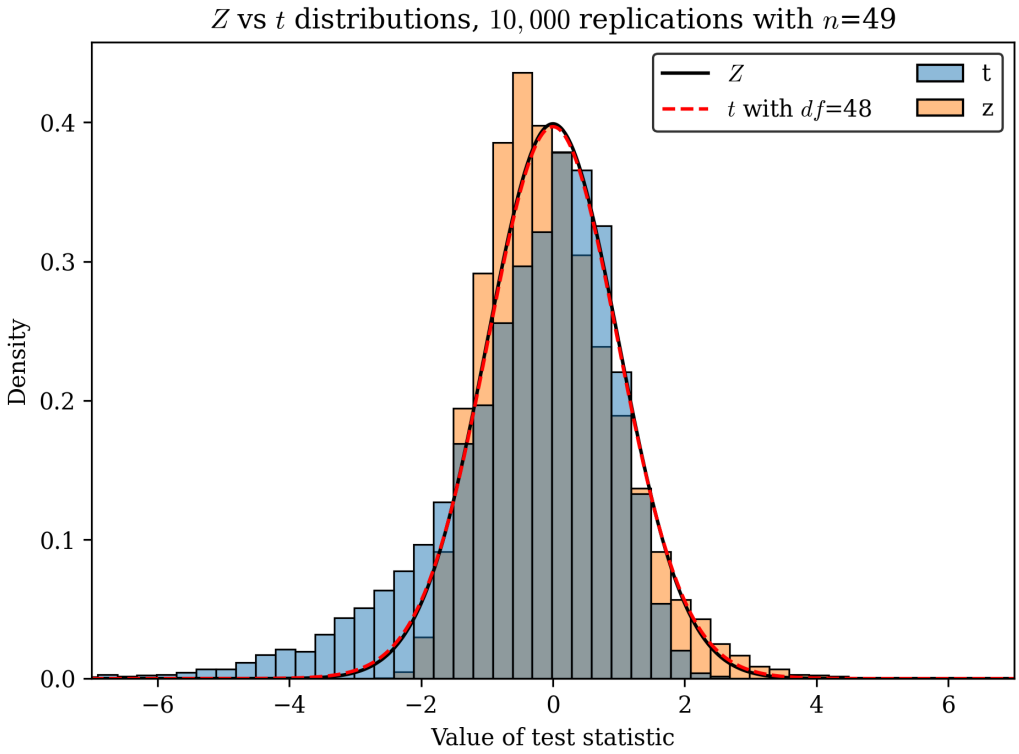
If that last paragraph was a bit heady, don't worry. You can just know that $df = n - n_k$, so that we get more degrees of freedom with larger sample sizes and fewer parameters that we need to estimate "along the way". Also, in the case of the t-distribution, $df > 100$ or so makes essentially *no* difference; in fact, using regular rounding procedures on the z-distribution. For example, we conventionally say that $z_{95} = 1.96$ but it is actually 1.959964...; if you have $df = 100,000$, which is possible with some samples like the CPS, your correct t-statistic is 1.959987.... In other words, with a large sample size, if you have a t-distributed variable and ignore this in favor of using a very exact z-score, that is a smaller mistake than in rounding z to two decimal places as almost everyone does.

2.2 Comparing t- and z-distributions

Here is a picture of how the t- and z-distributions relate with various df.



That's really all there is to it. Let's look at an example. Here is a sample of size $n = 49$ from the Titanic data-set. First, this is a case where we *do* know the sampling distribution because we know the population. So, let's simulate our sampling distribution first, just to get a look at it. We see that neither distribution is a *perfect* fit for its theoretical distribution (partially because of the small sample size, partially because of the finite if large number of replications). But, the t is substantially more skewed.

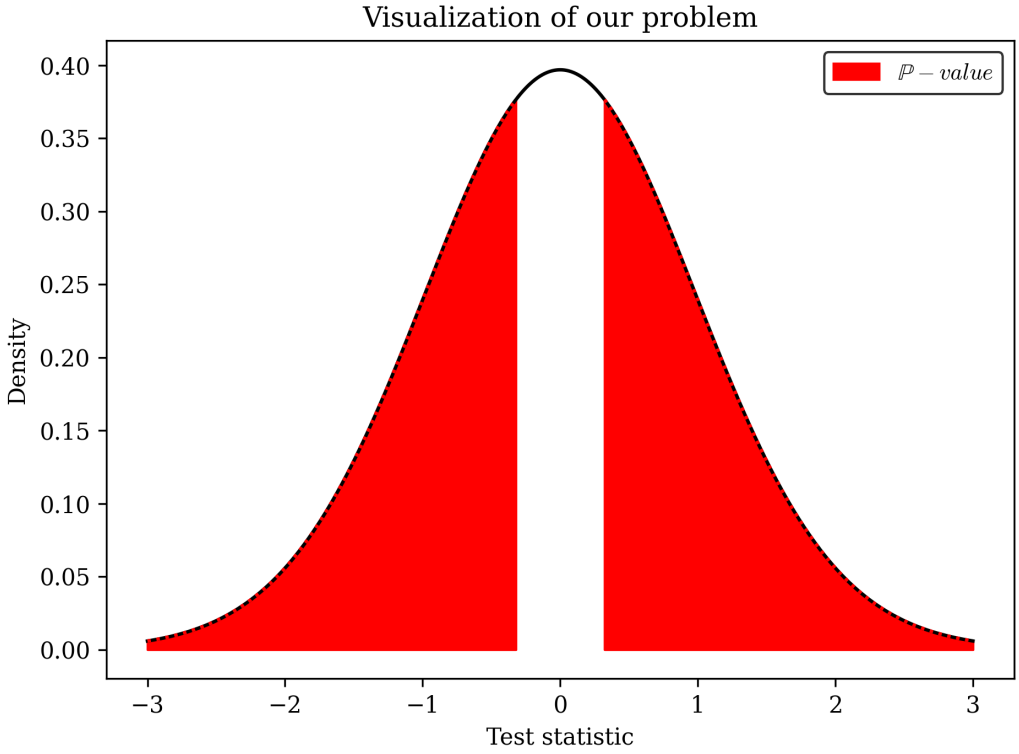


Now, let's examine the sample data and carry out some statistical tests.

	fare
count	49.00
mean	37.14
std	46.50
min	7.25
25%	10.50
50%	18.75
75%	46.90
max	262.38

First, test the claim that the mean fare in the population was 35. ***From here on out, we will generally focus on two-sided tests given the problems associated with one-sided tests that I previously mentioned, and we will generally prefer confidence intervals***. Then, find a 95 CI for the fare.

To test that claim, we find $t = \frac{37.14-35}{46.50/7} = 0.32$. Let's picture this situation, then solve.



OK, so, we know that we want to find the AUC to the right. So, using our *left-tail probability* in Stata, we need the complement with 1. We write `di 1-t(48, 0.32)` and get back 0.38; doubling, we have a whopping 0.76, that is, $\mathbb{P}$[data this extreme or more$|H_0$] is *very* large; we have no real evidence against the null. So, we fail to reject the null at any conventional α, such as 0.05. What if we'd done a one-tailed test in the right direction? Still no evidence at conventional levels, but we'd have $\mathbb{P} = 0.38$. If we'd gone in the wrong direction? We would just find `di t(48, 0.32)`, which we could also find by simply taking our first answer from 1: `di t(48, 0.32) = di  1-(1-t(48, 0.32))`, which both equal 0.68.

How about our CI? From here on out, I won't draw pictures of CIs because you don't need them nearly as badly, in part since there are infinite pictures you could draw of possible means μ which are compatible with our data (for some such pictures, see the original lecture slides on CIs).

1. Write out the formula for our specific case.

   Here, that's $CI_{95} = 37.14 + / - t_{95} \cdot SE$.

2. Solve for t_C, here with $C = 95$.

   Here, `di invt(48, 0.025)` $= 2.01$

3. Solve for the SE.

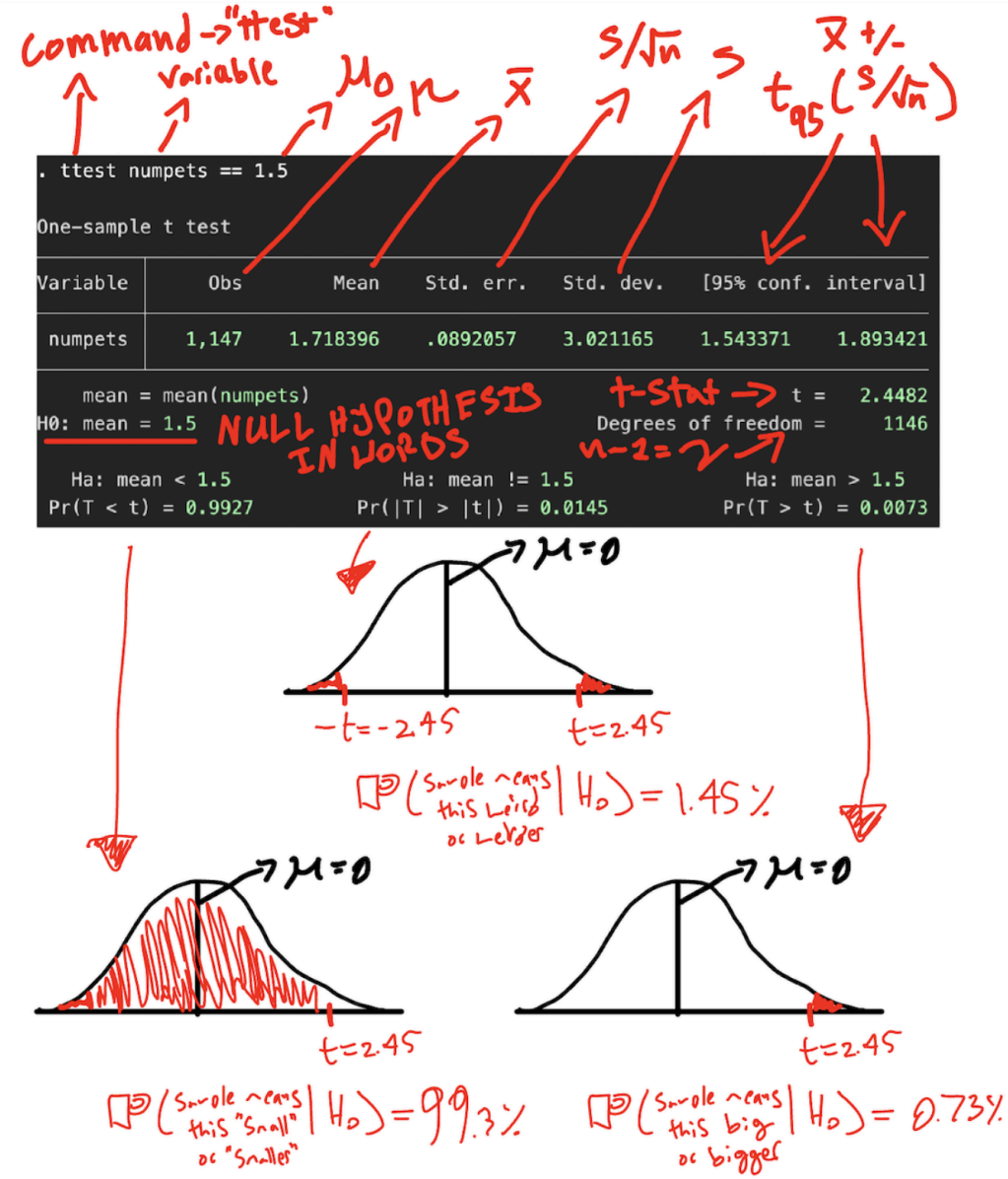
   Here, we simply have $\frac{46.50}{7}$, which we already calculated (it's 6.64).

4. Plug and chug.

   Here, that's $CI_{95} = 37.14 +/- 2.01 \cdot 6.64$.

   We get $LB = 23.79, UB = 50.49$.

How do we do this *automatically* in Stata? Pretty simple: `ttest $yvar$ == $your value$`. Here is annotated output for the `numpets` variable in the GSS we played around with last week.



3 The standard error for most of our remaining procedures all in one fell swoop: Bienaymé's identity

Now, let's generalize to cases where we have as our statistic of interest a difference between *two* population means, a single population proportion (or percent or probability), or a difference in proportions. **I'm going to go through the remainder of the material covered in Moore *et al.* quite quickly. Here is how and why.**

How? First, the authors correctly tell you that there is a common formula for *all* of the test-statistics and confidence intervals. It is this:

$$\text{test statistic} = \frac{\text{sample statistic} - \text{hypothetical parameter}}{\text{standard error}}$$
$$\text{confidence interval} = \text{sample statistic} +/- (\text{critical value} \cdot \text{standard error})$$

It is very easy to swap in and out the proper sample statistics and parameter values. The standard error formulae are a bit trickier, but the authors fail to tell you that there is a very simple pattern for all the different standard error formulae. We have already found the sample mean's standard error using a bit of variance algebra for sums of random variables (where we considered our sample mean to be the ratio of a sum to a constant, n). I'll remind you of the details in a moment.

It turns out that we can use that same trick again! If we then want the variance of a sum *of* sums, there is a simple formula for that, and it is actually the very same identity, called Bienaymé's identity. Then, we just realize that a difference in means is just a sum (divided by a constant) plus negative one times another sum. All of these constants are very simple to deal with in variance algebra.

3.1 Bienaymé's identity

Let's now revisit Bienaymé's identity, which we have already seen since *this is essentially just the proof of the variance of the sample mean*!

First, **the variance of a sum of random variables (or a difference; just put in a negative sign) is the sum of the individual variables' variances and all their covariances.** I'm booting the proof to the appendix, but I've shown it on the board a couple times. It's not hard, but let's just get to the result. The double sum just means "fix a value of j, then sum over value of k that isn't j". What happens when $j = k$? Then that's just a variance and it's already been counted. Note that $\mathbb{V}$ just means "variance" and is sometimes used in place of σ^2 where it might be clunky to have a lot of stuff in a subscript. $\sigma_{X,Y}$ means the covariance of X and Y.

$$\mathbb{V}\left\{\sum_{j=1}^{p} Y_j\right\} = \sum_{j=1}^{p} \sigma^2_{Y_j} + \sum_{k!=j}^{p}\sum_{j=1}^{p} \sigma_{Y_j,Y_k}$$

3.1.1 First use of the identity: standard error of the mean

Now, we can use this proof in *two* separate and useful ways. The first is to get the sampling variance of a mean (from which we easily find the standard error; just take the root).

First, if $p = n$, i.e. the size of a sample, and the variables are *independent* (the value of one variable does not affect another so there is no covariance) and *identically distributed* (random draws from the same population with replacement or without replacement but a very large population), we have the formula for the variance of a sample total on our hands: the variances are all just σ^2_Y and the covariances are all 0, thus the sum is $n\sigma^Y$. If we have a *mean*, recall that we divide the whole sum by a constant n and this factors out as a square (as does any constant in the variance formula, because of the squaring). One n cancels and we have that the sampling variance is $\frac{\sigma^2_Y}{n}$ (take the root to get the SE).

3.1.2 Second use of the identity: standard error of a difference in means

The next part is kind of meta. Stick with me. Now, the standard error formula we just calculated, $\frac{\sigma^2_Y}{n}$, is *itself* a variance, no? It's the variance of a mean across samples. That means that if we now want to find the variance of a difference in two random variables, the sample mean of group $a, \bar{Y}_a$ and the sample mean of group $b, \bar{Y}_b$, we just use the same formula!

Here's the kicker: sometimes these two groups will have a covariance between their means. When does this happen? When the two groups are somehow linked *individually*. Read that again. This does *not* mean "men and women generally have a social relation in the US". That is not *necessarily* individual. If you sample men and women without *individually linking them*, the random fluctuations in men's mean wage won't have anything to do with random fluctuations in women's mean wage, right?

However, what if you sample married couples? Then, obviously, if the male group has a higher mean wage, so will the female group, because the correlation between husbands' and wives' earnings is generally quite large.

How do we translate this into symbols?

$$\begin{aligned}
\mathbb{V}[\bar{Y}_a - \bar{Y}_b] &= \sum_{j=1}^{p} \sigma^2_{\overline{Y}_j} + \sum_{k!=j}^{p} \sum_{j=1}^{p} \sigma_{\overline{Y}_j, \overline{Y}_k} \\
&= \sum_{j=1}^{2} \sigma^2_{\overline{Y}_j} + \sum_{k!=j}^{2} \sum_{j=1}^{2} \sigma_{\overline{Y}_j, \overline{Y}_k} \\
&= \sigma^2_{\overline{Y}_a} + \sigma^2_{\overline{Y}_b} - 2\sigma_{\overline{Y}_a, \overline{Y}_b}
\end{aligned}$$

Now, how does this affect our t procedures? Well, what if our population parameter is a difference in population means and the corresponding sample statistic is a difference in sample means? Our population mean is *often, not always* zero. And our denominator should be the root of the variance of the sample difference. So, we have generically…(note that the covariance is subtracted because we're not summing but taking the difference of variables)

$$\text{two-sample test statistic} = \frac{(\bar{Y}_a - \bar{Y}_b) - (\mu_{Y_a} - \mu_{Y_b})}{\sqrt{\frac{\sigma_a^2}{n_a} + \frac{\sigma_b^2}{n_b} - 2\sigma_{\overline{Y}_a, \overline{Y}_b}}}$$

$$\text{two-sample CI} = (\bar{Y}_a - \bar{Y}_b) + / - t_C \sqrt{\frac{\sigma_a^2}{n_a} + \frac{\sigma_b^2}{n_b} - 2\sigma_{\overline{Y}_a, \overline{Y}_b}}$$

And in many cases—no matched pairs, null hypothesis of zero—this reduces to…

$$\text{two-sample test statistic} = \frac{\bar{Y}_a - \bar{Y}_b}{\sqrt{\frac{\sigma_a^2}{n_a} + \frac{\sigma_b^2}{n_b}}}$$

$$\text{two-sample CI} = (\bar{Y}_a - \bar{Y}_b) + / - t_C \sqrt{\frac{\sigma_a^2}{n_a} + \frac{\sigma_b^2}{n_b}}$$

Note in general that we we add *variances*, not standard deviations. Why? The same subject space picture as before! *The squared of length of a combination of vectors is the combination of their squared lengths; the same is* ***not*** *true of unsquared lengths.* This means that many standard presentations of confidence intervals, e.g. the ubiquitous "error bar" graph, should not be used *directly* to infer whether a difference in means overlaps.

What if we don't know the standard deviations? Once again, we just swap in the sample estimates. This actually causes the resulting distribution to be something *other than the* t, itself already a deviation from the Normal. However, this hardly matters, and most textbooks skip over this fact lightly (indeed, most textbooks should skip lightly over the t-distribution in the first place, but change is slow).

This then gives us the formulae for the actual t***-procedures.***

$$t_{\text{two-sample}} = \frac{(\bar{Y}_a - \bar{Y}_b) - (\mu_{Y_a} - \mu_{Y_b})}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b} - 2\hat{\sigma}_{\overline{Y}_a, \overline{Y}_b}}}$$

$$t_{\text{two-sample, most of the time}} = \frac{\bar{Y}_a - \bar{Y}_b}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}}$$

$$\text{CI}_{\text{two-sample}} = (\bar{Y}_a - \bar{Y}_b) + / - t_C \sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b} - 2\hat{\sigma}_{\overline{Y}_a, \overline{Y}_b}}$$

$$\text{CI}_{\text{two-sample, most of the time}} = (\bar{Y}_a - \bar{Y}_b) + / - t_C \sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}$$

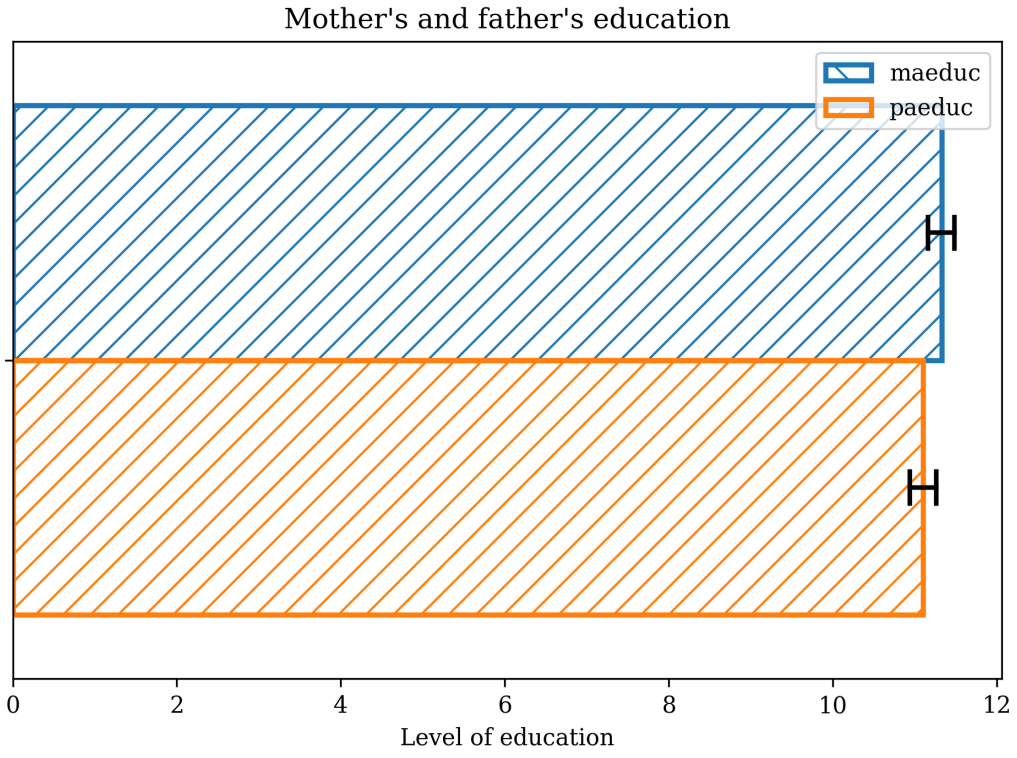
3.2 Seeing why it matters: two-group test procedures

We can then come up with formulae for our remaining procedures very quickly. But first, what difference does all of this make? Generally, it is important to have *correct* standard errors for obvious reasons. But many times getting things correct also specifically *makes our estimates more precise in systematic ways.*

Here's a simple example. Perhaps you're interested in whether men's and women's years of education differ and you have data on people who were married or at least had a kid together (you can get this information from the GSS, which asks respondents about their parents).

3.2.1 A common mistake to motivate us

You'll often see data represented in bar graphs like the one below out there in the wild. Here we have data on mother's and father's education levels for working class respondents to the GSS in 2018. I've restricted this to parents of working class (self-identified) children to get numbers that prove my point. These are really matched pairs (why? We don't just have random groups of men and women), but this standard error approach is actually misleading.



This figure seems to imply that the two groups do not have different means because their error bars overlap. *Note that in this case, as is often the case, the error bars are just the mean* $+/-$ *one standard error,* i.e. *the represent a* 68 *percent confidence interval*, or something very close to it when n is bigger than about 100 (here, both groups have $n = 635$). That

level of confidence is arbitrary, but it is often used because it makes the math really simply.

However, at a 68 percent level of confidence, *do* they differ? First, let's think about what formula we are *implicitly* using in comparing the error bars. We are *implicitly* looking for a case where the upper bound of group b exceeds the lower bound of group a, which we can rearrange…

$$\begin{aligned} \bar{Y}_a - SE_a &< \bar{Y}_b + SE_b \\ \bar{Y}_a - \bar{Y}_b - (SE_a + SE_b) &< 0 \end{aligned}$$

Aha! This now looks like a 68 percent confidence interval, only it's *got the wrong standard error*. We could calculate this and see, indeed, that this pseudo-CI includes zero (so there appears to be no difference). The we can use the *correct* formula. Here are the data below. Let $a =$ women, $b =$ men.

	maeduc	paeduc
count	635.00	635.00
mean	11.33	11.10
std	3.87	4.16
min	0.00	0.00
25%	10.00	9.00
50%	12.00	12.00
75%	14.00	13.00
max	20.00	20.00

First, the wrong formula:

$$\begin{aligned} \text{pseudo-lowerbound} = \bar{Y}_a - \bar{Y}_b - (SE_a + SE_b) &< 0 \\ (11.33 - 11.1) - \left\{ \frac{3.87}{\sqrt{635}} + \frac{4.16}{\sqrt{635}} \right\} &= -0.09 \end{aligned}$$

Since zero is larger than the lower bound, we can't confidently reject the null of no difference between the two groups.

3.2.2 The regular formula for a standard error for the difference in means (unknown σ_a, σ_b)

Now, let's execute the two-group formula. This still fails to take advantage of the matching, but it's closer to the truth.

$$
\begin{aligned}
CI_{\text{no matching}} &= (\bar{Y}_a - \bar{Y}_b) + / - t_C \sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}} \\
&= (11.33 - 11.1) + / - 1 \cdot \sqrt{\frac{3.87^2}{635} + \frac{4.16^2}{635}} \\
&= (11.33 - 11.1) + / - 0.2255 \\
&= 0.005, 0.456
\end{aligned}
$$

So, this is *very* close, but zero does *not* appear in the confidence interval here, so we *do* have evidence of a difference in means, favoring women, at this level of confidence.

By the way, you can do this in Stata, but since the data really are paired (in Stata, this is equivalent to the fact that they are separate variables, not one variable which can be separated by values of another), we need to do so semi-manually. Here, execute `ttesti 635 11.32598 3.866364 635 11.10079 4.163752, level(68) unequal` (your answer will differ slightly because of rounding). The generic syntax for this is `ttesti n_a mean_a std_a n_b mean_b std_b, unequal`. The "unequal" option is actually what should be standard, since the default assumes equal variance in groups; as Moore *et al.* point out, that's rarely a good assumption.

How do we do this test *automatically* in Stata if we don't have matched pairs? Simple. `ttest $outcome$, by($predictor$) level($confidence as an integer$) unequal`. Try this out with, say, `educ` and `sex`. ***Ignore the "combined" row and focus on the difference row and the material below that, which doesn't change in interpretation from the one-sample case.***

3.2.3 The regular formula for a standard error for the difference in means of matched pairs (unknown σ_a, σ_b)

However, the *even more correct* approach here is to treat these as matched pairs since they are. Why do this? Well, we get to subtract out the sampling covariance between the two means. Fortunately, you don't even have to do this manually (although the sampling covariance isn't hard to find; see the appendix). You can just make a *difference variable* and conduct one-sample t-procedures, or even turn it over to Stata directly with `ttest maeduc == paeduc if class==2`.[1] Both of these directly estimate the formula with the covariance subtracted out.

When we do that, we get $0.1, 0.35$. So, at a 68 percent level of confidence, we are even farther from seeing zero in the interval; in other words, we're pretty sure that there is some difference at this small level of confidence! Of course, the level of confidence is ultimately arbitrary, but ***the key point is that using matched pairs increases the power of our test by shrinking the standard error through no trickery.***

[1] Common coding mistake is to only use single equals signs, but in code this is always an assignment, something that you *make* true by definition. To check for equality in a case where it *might* be false, you always need double equals signs, `==`.

3.2.4 Two-group degrees of freedom and tests

By the way, we only used confidence intervals above, and I punted a little bit on how exactly to find degrees of freedom (we just used $t_{68} \approx 1$). That's because you should mostly do this in Stata. Differences in means are actually not even quite t-distributed, only close, and degrees of freedom for this are tricky and can be non-integers.

If you're working by hand and don't have matched pairs, you should just use the smaller n of your two groups. If you *do* have matched pairs, you only have one group, really, and you should just work in Stata (you can't realistically make a difference variable $y_a - y_b$ by hand with large n). This also frees you of the need to actually find the intuitive but fairly involved formula for the SE of the difference in correlated sample means. Making a difference vairable and doing a one-sample test *or* just writing `ttest mean_a == mean_b` will automatically estimate that difference.

So, if we weren't using a matched pairs design, to find a 95 percent CI , we would use $n = 635$ (here, our two groups have the same size anyways). We can write in Stata `di invt(635, 0.025)` and we get 1.96, just as we would with a z (like I said, t-distributions are basically z-distributions for $n > 100$).

Then, our 95 CI is simply $0.23 - 1.96(0.225), 0.23 + 1.96(0.225) = -0.21, 0.67$. Note that we could also have just taken our early `ttesti` code and just swapped in `level(95)`, and the CI is in the output. We once again have no evidence of a difference in means, but that's because we changed our level of confidence.

To conduct a test, we can simply find $t = \frac{0.23}{0.255}$ and then `di 1-t(635, 0.23/0.225)` to get a right-tail probabilty, and then double it. We get about 0.31 here. Generally, I'm OK with you doing these things in Stata.

3.2.5 Non-zero null in Stata

An annoying limitation of Stata is that it is generally hard to use a non-zero null. There is no default option, actually. Here is a quick workaround.

1. Decide which group you think has a higher mean.

   Suppose that we want to test whether verbal acuity differs between generations. Let's make a dummy. This is good review.

   ```
   use gss2018, clear
   gen fortyplus = age > 39
   replace fortyplus = . if missing(age)
   lab def fp 0 "under 40" 1 "40 or older"
   lab val fortyplus fp
   tab fortyplus, sum(wordsum)
   ```

   Preliminary inspection reveals that older people have a higher mean. Let's test whether it was as large as a half-point gap (5 percent since it's a ten-item test).

2. Decide how much of a difference you want to test; add this amount to the lower-scoring group. Testing whether the two groups are equal *on this new variable* is equivalent to testing whether their score differs by this amount.

Here, we'll make a new variable called `testwordsum` that gives the younger people half a point. *Note that you can get confusing results if the difference is too large, so make sure you pick an amount that is smaller than the observed gap.*

```
gen testwordsum = wordsum
replace testwordsum = wordsum + 0.5 if fortyplus==0
```

3. Conduct a regular t-test in Stata.

   ```
   ttest testwordsum, by(fortyplus) unequal
   ```

   Here, we have no evidence of a difference. We can manually confirm. Below are the data.

fortyplus	count	mean	std	min	25%	50%	75%	max
False	509.00	5.64	1.96	0.00	4.00	6.00	7.00	10.00
True	1031.00	6.05	2.04	0.00	5.00	6.00	7.00	10.00

$$\begin{aligned} t &= \frac{(\bar{y}_a - \bar{y}_b) - (\mu_{a0} - \mu_{b0})}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}} \\ &= \frac{(6.05 - 5.64) - 0.5}{\sqrt{\frac{2.04^2}{1031} + \frac{1.96^2}{509}}} \\ &= 0.85 \end{aligned}$$

We could find our $\mathbb{P}$-value here, but we've done enough to confirm.

3.2.6 Two-group tests can be done with regression

We saw a couple weeks ago that you can also test differences in groups by running a regression. Sticking with our example above, try `reg wordsum fortyplus`. Note that the t-statistic on the slope is the same t-statistic that you get from `ttest wordsum, by(fortyplus) unequal`. Remember that the constant is the mean for the 0-coded group and the slope is the difference in means.

3.3 The remaining formulae: one group proportion, difference in proportions

Now we can quickly develop our proportion procedures. What is even new here at all? After all, we previously knew about binomially-distributed sample *counts* T. We could already do hypothesis tests about p using sample counts, but it is somewhat more natural to do this with proportions, and confidence intervals are even a bit more natural. **The main reason for the difference is that a proportion can take on non-integer values, so it is a continuous variable, just like before when we first learned z-procedures**.

Our basic structure does not change.

$$\text{test statistic} = \frac{\text{sample statistic} - \text{hypothetical parameter}}{\text{standard error}}$$
$$\text{confidence interval} = \text{sample statistic} +/- (\text{critical value} \cdot \text{standard error})$$

But, sample proportions are both estimates of the population probability of a dummy *and* sample means. So, the standard error of a proportion is just the standard error of the sample total with an $\frac{1}{n^2}$ factored out. The standard error of the total we already knew as $\sqrt{np(1-p)}$, so now we just have $\sqrt{\frac{p(1-p)}{n}}$ as our standard error of a proportion. Handy! *Even better, because proportions' variances are fully determined by their mean, the sample statistic is a z-score, so we can just use the Normal.*

By the way, in what follows, I just show the formulae for sample estimates of the standard errors because they differ based on whether we're doing a test or a CI, so the abstract form assuming we know the population is just one thing too many.

3.3.1 The z-statistic for a single-sample proportion

So, we have for a single sample something that should look very familiar. Here, $\hat{p}$ means the sample proportion or estimated population probability (same thing). Note that in general a "hat" means "a sample estimate". We don't always use this notation—e.g., $\bar{y}$ should really be $\hat{\mu}_Y$. I thought about bucking convention and just teaching with this consistent, clear notation, but I didn't want to contradict established usage too much. p_0 means, as always, our hypothetical population probability.

$$\text{test statistic} = \frac{\text{sample statistic} - \text{hypothetical parameter}}{\text{standard error}}$$
$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Why the p_0 in the denominator? Well, we assume that the null is true, so the sampling distribution *under the null* has a variance given by the hypothetical mean, not the real mean.

Here's a quick example. Suppose we want to test the claim that the population proportion of people who are divorced is 0.15, using the GSS. Let's just use a two-tailed test.

divorced	count	share
False	1943	0.83
True	403	0.17
total	2346	1

Let's now apply our formula.

$$
\begin{aligned}
z &= \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \\
z &= \frac{0.17 - 0.15}{\sqrt{\frac{0.15(1-0.15)}{2346}}} \\
z &= 2.95
\end{aligned}
$$

Then, `di 2*(1-norm(z))` $= 0.003$. Very strong evidence against the null! You can check your work in Stata with the following code.

```
d marital
lab li MARITAL // OK, divorced is "3"
gen divorced = marital == 3
replace divorced = . if missing(marital)
    // remember, we do this because one way
    // for the condition "marital == 3" to be
    // false is for it to be missing, but
    // we don't want those people assigned
    // to the condition "0" on divorced.
prtest divorced == 0.15
```

3.3.2 The CI for a single-sample proportion

Next, we have the formula for the CI.

$$\text{confidence interval} = \text{sample statistic} +/- (\text{critical value} \cdot \text{standard error})$$

$$CI = \hat{p} +/- z_C \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Why do we use $\hat{p}$ here? We have no null, so we have no particular value to use.

Let's do a somewhat dry example and get a 95 CI for the proportion who are divorced. Here, it is not totally redundant since we do need to use the *observed*, not some hypothesized, proportion in calculating the CI.

$$
\begin{aligned}
CI &= \hat{p} +/- z_C \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \\
&= 0.17 +/- 1.96 \cdot \sqrt{\frac{0.17 \cdot 0.83}{2346}} \\
&= 0.1547, 0.1853
\end{aligned}
$$

You can check your work with `ci prop divorced, wald`. The "Wald" option (named for the great statistician, Abraham Wald; the famous "survivor bias plane" diagram is based on a paper he wrote) ensures that Stata uses the formula we use.

3.3.3 The CI for a difference in proportions between two groups

Finally, let's get our two-sample procedures going. First, the z confidence interval. Note that although this looks very complicated, it obeys a simple rule. First, we have our point estimate or our sample statistic—here a difference in proportions—plus and minus our standard error. This looks difficult, but keep in mind that it is easy to build up piece-by-piece. First, we reasoned above that our difference in sample proportions is just another case of a difference in two random variables, and the variance of a sum or a difference of two *independent* random variables is given by the sum of their variances (they have no covaraince). Here, the variance of each random variable is *also* covered by the same formula for the variance of a sum because a mean is just a "deflated" sum, so the variance of each mean is just $\frac{\sigma^2}{n}$. For a dummy, $\sigma^2 = p(1-p)$, and each group has its own $n : n_a, n_b$.

$$CI_{\text{two-sample prop.}} = (\hat{p}_a - \hat{p}_b) +/- z_C \cdot \sqrt{\frac{\hat{p}_a(1-\hat{p}_a)}{n_a} + \frac{\hat{p}_b(1-\hat{p}_b)}{n_b}}$$

Here's an example. Let's say that our two groups are college graduates and non-college-graduates. Let's see if the groups differ in their opinion on whether abortion should be allowed under all circumstances. Conveniently, this is already a dummy variable in the GSS (although not coded correctly; more in a moment).

collgrad	count	mean	var
False	930	0.44	0.25
True	594	0.59	0.24
total	1524	0.5	0.25

What is our 95 CI for the difference?

$$\begin{aligned} CI_{\text{two-sample prop.}} &= (\hat{p}_a - \hat{p}_b) +/- z_C \cdot \sqrt{\frac{\hat{p}_a(1-\hat{p}_a)}{n_a} + \frac{\hat{p}_b(1-\hat{p}_b)}{n_b}} \\ &= (0.59 - 0.44) +/- 1.96 \cdot \sqrt{\frac{0.25}{928} + \frac{0.24}{594}} \\ &= (0.1016, 0.2031) \end{aligned}$$

Here's how to do this in Stata.

```
use gss2018, clear
tab abany, gen(abdum)
rename abdum1 abort_any_reason
replace abort_any_reason = . if missing(abany)
lab def abreas 0 "no" 1 "yes"
lab val abort_any_reason genericlab
d deg
```

```
lab li LABAM
gen collgrad = degree > 1
replace collgrad = . if missing(degree)
lab def cg 0 "not a grad" 1 "is a grad"
lab val collgrad cg
prtest abort_any_reason, by(collgrad)
```

3.3.4 The z-statistic for a difference in proportions between two groups

And finally the test-statistic. Why has the standard error changed yet again? Once more, consider that under the null hypothesis, the two groups don't have a difference. **We're not actually forced to make a hypothesis about what their pooled proportion *is*—**we're interested in whether they differ—***so here we just estimate the pooled proportion from the data*.

For non-zero nulls, things get more complicated, unlike two-sample t-tests, again because the variance of a proportion is fully determined by the proportion: there are many ways that two groups' proportion of success on an outcome could differ by a fixed percentage, so we will just generally suppose $H_0 : p_a - p_b = 0$.

$$\begin{aligned} z_{\text{two-sample prop.}} &= \frac{(\hat{p}_a - \hat{p}_b) - (p_{a0} - p_{b0})}{\sqrt{\text{sampling variance of group } a + \text{sampling variance of group } b}} \\ &= \frac{(\hat{p}_a - \hat{p}_b)}{\sqrt{\hat{p}(1-\hat{p}) \cdot (\frac{1}{n_a} + \frac{1}{n_b})}} \end{aligned}$$

How's this work out in the case of our example?

$$\begin{aligned} z_{\text{two-sample prop.}} &= \frac{(\hat{p}_a - \hat{p}_b)}{\sqrt{\hat{p}(1-\hat{p}) \cdot (\frac{1}{n_a} + \frac{1}{n_b})}} \\ &= \frac{(0.59 - 0.44)}{\sqrt{0.25 \cdot (\frac{1}{594} + \frac{1}{930})}} \\ &= 5.80 \end{aligned}$$

We could convert this into a $\mathbb{P}$-value, but you hopefully know by now that this is assuredly statistically significant at any conventional level.

3.3.5 Comment on the complexity of the formulae

It's more important to know how the formulae *work*, but you'll rarely need to do this by hand. ***On the homework and the final, I will ask you some conceptual questions about these formulae, but I won't ask you to do extensive hand-calculation***. As long as you follow why the formulae are what they are above, I am licensing you to just use

Stata. The main thing you need to know is that your variables must not only be binary but also 0/1-coded, which many variables are not unless we've made them that way (as we did above).

There are many ways to solve this. A very quick one to try is `tab $yourbinary var$, gen($var$dum)`. This creates a set of dummy variables corresponding to every possible value of a variable; if your variable was already a binary categorical variable, this just makes two 0/1 binary variables for each level. However, you should still rename and relabel your variable for clarity.

3.4 Standard error for the bivariate regression slope

How about inference on the bivariate regression slope? We can do that *parametrically*, as our last exercise of this sort, before we look at the bootstrap. Here, I'm following Cosma Shalizi's proof. I'll follow his usage in *not* using the $n-1$ correction.

First, we write the sample regression slope as follows. Note that ϵ_j is the *population* error, not the sample error, so its mean in the sample is not necessarily zero.

$$
\begin{aligned}
\hat{\beta}_1 &= \frac{\hat{s}_{X,Y}}{s_X^2} \\
&= \frac{\frac{1}{n}\sum_{j=1}^{n} x_j y_j - \bar{x}\bar{y}}{s_X^2} \\
&= \frac{\frac{1}{n}\sum_{j=1}^{n} x_j(\beta_0 + \beta_1 x_j + \epsilon_j) - \bar{x}(\beta_0 + \beta_1 \bar{x} + \bar{\epsilon})}{s_X^2} \\
&= \frac{\beta_0 \bar{x} + \beta_1 \overline{x^2} + \frac{1}{n}\sum_{j=1}^{n} x_j \epsilon_j - \bar{x}\beta_0 - \beta_1 \bar{x}^2 - \bar{x}\bar{\epsilon}}{s_X^2} \\
&= \frac{\beta_1 s_x^2 + \frac{1}{n}\sum_{j=1}^{n} x_j \epsilon_j - \bar{x}\bar{\epsilon}}{s_X^2} \\
&= \beta_1 + \frac{\frac{1}{n}\sum_{j=1}^{n} x_j \epsilon_j - \bar{x}\bar{\epsilon}}{s_X^2} \\
&= \beta_1 + \frac{\frac{1}{n}\sum_{j=1}^{n} (x_j - \bar{x})\epsilon_j}{s_X^2}
\end{aligned}
$$

Now, if we take the variance of the variable, we can treat the x_js as temporarily fixed (technically appropriate only in designed experiments, but we'll get around that in a moment).

$$\begin{aligned}
\mathbb{V}[\hat{\beta}_1] &= \mathbb{V}\left\{\beta_1 + \frac{\frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})\epsilon_j}{s_X^2}\right\} \\
&= \left\{\frac{\frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})}{s_X^2}\right\}^2 \mathbb{V}[\epsilon_j]
\end{aligned}$$

Then, assuming that the ϵ_js are IID random variables, we simply have (letting σ with no subscript refer simply to the conditional standard deviation)…

$$\begin{aligned}
\mathbb{V}[\hat{\beta}_1] &= \left\{\frac{\frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})}{s_X^2}\right\}^2 \sigma \\
&= \left\{\frac{\frac{1}{n^2}\sum_{j=1}^{n}(x_j - \bar{x})^2}{(s_X^2)^2}\right\}\sigma \\
&= \frac{\sigma^2}{n s_X^2}
\end{aligned}$$

Now, we assumed that the x_js were non-random. However, they are usually are functions of the sampling process, of course (not always, though: think experiments, which we see time-and-again are the simplest kind of statistical process—we can get away with simple linear regression, of which ANOVA or even a simple two-group test is a subspecies).

So, let's use the law of total variance. Recall that this works like the following…

$$\begin{aligned}
\mathbb{V}[Y] &:= \mathbb{E}[(Y - \mu_Y)^2] \\
&= \mathbb{E}[(Y - \mu_{Y|X} + \mu_{Y|X} - \mu_Y)^2] \\
&= \mathbb{E}[(Y - \mu_{Y|X})^2] + \mathbb{E}[\mu_{Y|X} - \mu_Y)^2] - 2\mathbb{E}[(Y - \mu_{Y|X})(\mu_{Y|X} - \mu_Y)] \\
&= \mathbb{E}_X\mathbb{E}_Y[(Y - \mu_{Y|X})^2|X] + \mathbb{E}_X\mathbb{E}_Y[\mu_{Y|X} - \mu_Y)^2|X] - 2\mathbb{E}_X\mathbb{E}_Y[(Y - \mu_{Y|X})(\mu_{Y|X} - \mu_Y)|X]
\end{aligned}$$

Now, the final cross-term drops out because the second term, $(\mu_{Y|X} - \mu_Y)$, does not depend directly on Y, so it can be moved outside of the expectation over Y. Then, the expectation over Y of $(Y - \mu_{Y|X}|X$ is always equal to zero (proven much earlier in this set of notes).

Then, we simply reinterpret the remaining quantities. The first term is the expectation over X of the conditional variance of Y, so it is the average conditional variance. The second term is the expectation of the squared deviation of the conditional means from the overall mean (note that the expectation over Y is actually unnecessary), so it is the variance of the conditional means.

We can write the law of total variance then as…

$$= \mathbb{E}_X \mathbb{V}[Y|X] + \mathbb{V}[\mu_{Y|X}]$$

Note that X can be a vector if we like, e.g.

We can write the law of total variance then as…

$$= \mathbb{E}_X \mathbb{V}[Y|X_1, X_2, ...X_n] + \mathbb{V}[\mu_{Y|X_1,X_2,...X_n}]$$

Then, finally, we have…

$$= \mathbb{E}_X \mathbb{V}[\hat{\beta}_1|X_1, X_2, ...X_n] + \mathbb{V}[\mu_{\hat{B}_1|X_1,X_2,...X_n}]$$

Since the mean of the sampling distribution doesn't depend directly on any particular sample, the second term is just the variance of a constant and drops out; then, the mean of the variance of $\hat{\beta}_1$ across values of the sample realizations of X is just the same quantity that we found above.

4 Bootstrapping

4.1 Parametric formulae

OK, if you're like me, you're probably a little sick of all of these *parametric* formulae. *Parametric* means that we are assuming something is true about the population. Of course, we are assuming certain null hypothesis values, but we're also assuming something about the *population sampling distribution*: that is Normal or t-distributed, that its variance is a fairly clean but often quite lengthy formula, and so on.

The types of statistical formulae above tend to be what make smart people bored with statistics, and not without reason. Their derivations are often beautiful and solved certain problems that were intractable without computers. However, digital computers are fantastically useful devices, and some of what we've learned above was designed for students in roughly the 1970s or earlier (not kidding). Here's another *major* problem with the above: it's all about means! We said nothing about medians/quantiles, or correlation coefficients, or the mode, or ratios of variables (beyond my point that they are hard to work out sampling variances for!). These have sampling distributions that are much more complicated.

4.2 Bootstrap: name and how it works

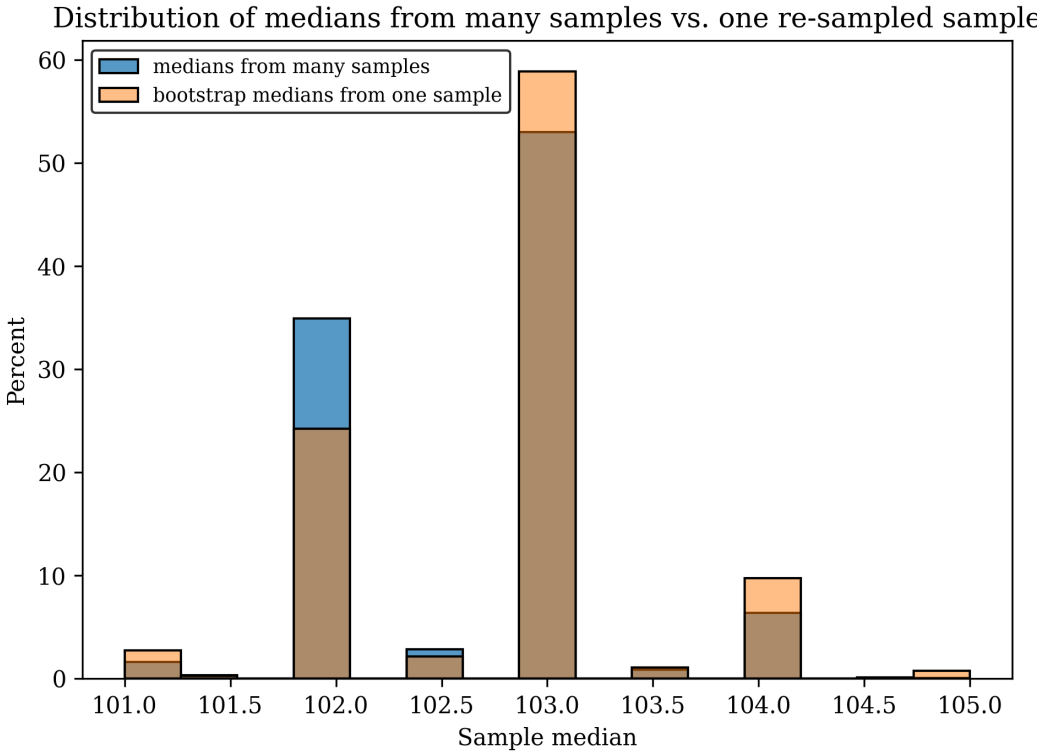
The "bootstrap" is a funny name for how we might solve some of these problems. The idea is the sample "pulls itself up by its own bootstraps" (you sociologists out there probably have heard this term used mockingly before; it's a bit of a cliché about how right-libertarians view the world). In fact, I have shown you boostrap results when I've conducted "simulations" for the entire semester!

OK, so what are the details? Well, basically, the idea is that if we have the ability to very easily resample our data-set, we can actually just *simulate* the sampling distribution by resampling our data over and over again.

We aren't actually "re-using" our data, a common misconception here, in a way that might seem to exhaust them. It's actually like this: we take our sample as a pretty good guide to the population probability of Y taking on this or that value. If we knew those values with *certainty*, we could get a bunch of samples *with a computer* that would hardly differ at all from the tricky business of going out and talking to people (I know, that sounds weird and antisocial, but sampling well is **very, very** hard). Nevermind why you would want to sample if you knew the population: the point is that if we knew the population distribution function, we could easily use computers to get samples that would not differ *statistically* from real world samples if we took many of both. So, we just hope that the sample we have is a pretty good guide.

The idea is that in a sample of, say, $n = 2000$, we'll resample all $n = 2000$ people ***with replacement***. This ensures that we get a sample that is as representative as possible of the population, assuming our sample is an OK estimate of the population distribution. If we sampled without replacement, we would need much smaller sample sizes and we would, at some point, have a much higher conditional probability of including extreme values that happened to end up in our sample; you can think about this as being *like* taking a sample mean vs. an individual observation: we reduce the influence of outliers in the population who might just end up in our sample.

Here is a picture of how that works using the NBA population data. What's happening here is a little tricky because this is a situation where we *know* the population. First, I pulled one sample of $n = 1000$ games and plotted each team's points scored from the population data. Then, I **resampled the sample with replacement** 5000 **times** (AKA bootstrapped it), with $n = 1000$. Then, I pulled 5000 *fresh* samples from the actual population and plotted those. The point here is that these distributions actually look quite similar. Why does this matter? ***It means that we can get an estimate of the standard error of the median by simply bootstrapping our sample***.



So what do we do with this set of statistics from many re-samples? We don't actually take the mean of the statistics, another common misconception. The expected value of our resamples is just the sample mean that we started with, which won't get us any closer to the true mean.

The point is actually to just get an "empirical" CI, which we do by simply finding the 2.5th and 97.5th percentiles (or whatever other ones you might like to get a C CI). This is called the *percentile bootstrap*, and we use it in place of the regular formula for the CI because the distribution of our sample statistic across many samples may not be Normal or anything close to it. *Since our re-sampling distribution is pretty close to our sampling distribution, the idea is that we can non-parametrically find the lower and upper bounds of the distribution.*

In this case, the population median was 103, and the values for the 2.5th and 97.5th percentiles of our resampled medians were $101, 104$, so our procedure worked!

4.3 When does bootstrapping work to generate a plausible confidence interval?

Almost always. You can even use it for means, whose CIs we know how to find with parametric formulae. You can also use it on statistics such as the regression slope, whose sampling distribution's variance is not hard to derive but is fairly complex. ***Standard errors are, in one sense, fully one-half of statistics: how much variation is there across the whole sampling process?*** So, we have to spend a lot of time on them,

and it's nice when we can simplify the process.

4.4 Bootstrapping example

Fortunately, this is also really easy to do in Stata. Below follows syntax for means, medians and regression slopes. You generally write `bootstrap $statistics you want to bootstrap$ reps($number here$): $the regular command$`. The default is to give you estimates based on the assumption that the sampling distribution is Normal; you can get percentile estimates by running the post-estimation command `estat bootstrap, all`. If there is no ambiguity about what the regular command might return that you'd want to bootstrap, you can omit that (as in regression). I threw in the regular CI for the mean as well for your comparison.

```
use gss2018, clear
bootstrap r(mean) r(p50), reps(1000) seed(1992): sum age, d
estat bootstrap, all
ci mean age
bootstrap, reps(1000): reg educ paeduc
estat bootstrap, all
```

5 Appendix

5.1 Bienaymé's identity

$$
\begin{aligned}
\mathbb{V}\left\{\sum_{j=1}^{p} Y_j\right\} &= \mathbb{E}\left\{\sum_{j=1}^{p} Y_j - \sum_{j=1}^{p} \mu_j\right\}^2 && \text{Read as the expectation of the square} \\
&= \mathbb{E}\left\{\sum_{j=1}^{p} (Y_j - \mu_j)\right\}^2 && \text{Summation simplification} \\
&= \mathbb{E}\left\{\sum_{j=1}^{p} (Y_j - \mu_j)^2 + \sum_{k\neq j}^{p}\sum_{j=1}^{p} (Y_j - \mu_j)(Y_k - \mu_k)\right\} && \text{Polynomial expansion} \\
&= \mathbb{E}\left\{\sum_{j=1}^{p} (Y_j - \mu_j)^2\right\} + \mathbb{E}\left\{\sum_{k\neq j}^{p}\sum_{j=1}^{p} (Y_j - \mu_j)(Y_k - \mu_k)\right\} && \text{Expectation is additive} \\
&= \sum_{j=1}^{p} \mathbb{E}[(Y_j - \mu_j)^2] + \sum_{k\neq j}^{p}\sum_{j=1}^{p} \mathbb{E}[(Y_j - \mu_j)(Y_k - \mu_k)] && \text{Again, expectation is additive} \\
&= \sum_{j=1}^{p} \mathbb{V}[Y_j] + \sum_{k\neq j}^{p}\sum_{j=1}^{p} \mathbb{COV}[Y_j, Y_k] && \text{Definition of variance and covariance} \\
&= \sum_{j=1}^{p} \mathbb{V}[Y_j] + 2\sum_{k>j}^{p}\sum_{j=1}^{p} \mathbb{COV}[Y_j, Y_k] && \text{Simplifying a summation which has many rep} \\
&= \sum_{j=1}^{p} \sigma^2_{Y_j} + 2\sum_{k>j}^{p}\sum_{j=1}^{p} \sigma_{Y_j, Y_k} && \text{Notational simplification}
\end{aligned}
$$

5.2 The sampling covariance of the mean.

First, here's how the covariance of a sum works generally.

$$\begin{aligned}
\mathbb{COV}\left\{\sum_{j=1}^{n} Y_{aj}, \sum_{k=1}^{n} Y_{bk}\right\} &= \mathbb{E}\left\{[\sum_{j=1}^{n} Y_{aj} - \sum_{j=1}^{n} \mu_a][\sum_{k=1}^{n} Y_{bk} - \sum_{k=1}^{n} \mu_b]\right\} \\
&= \mathbb{E}\left\{[\sum_{j=1}^{n} (Y_{aj} - \mu_a)][\sum_{k=1}^{n} (Y_{bk} - \mu_b)]\right\} \\
&= \mathbb{E}\left\{\sum_{k=1}^{n}\sum_{j=1}^{n} (Y_{aj} - \mu_a)(Y_{bk} - \mu_b)\right\} \\
&= \sum_{k=1}^{n}\sum_{j=1}^{n} \mathbb{E}[(Y_{aj} - \mu_a)(Y_{bk} - \mu_b)] \\
&= \sum_{k=1}^{n}\sum_{j=1}^{n} \mathbb{COV}[Y_{aj}, Y_{bk}]
\end{aligned}$$

From here, we can write our covariance of means as

$$\mathbb{COV}[\bar{Y}_a, \bar{Y}_b] = \mathbb{COV}\left\{\frac{1}{n}\sum_{j=1}^{n} Y_{aj}, \frac{1}{n}\sum_{k=1}^{n} Y_{bk}\right\}$$

Then, we just pull out an n^2.

$$\frac{1}{n^2}\mathbb{COV}\left\{\sum_{j=1}^{n} Y_{aj}, \sum_{k=1}^{n} Y_{bk}\right\}$$

Then use the conclusion from above.

$$= \frac{1}{n^2}\sum_{k=1}^{n}\sum_{j=1}^{n} \mathbb{COV}[Y_{aj}, Y_{bk}]$$

And finally, we note that since the draws from our random variable are independent *except* within matched pairs, we have simply

$$= \frac{1}{n^2}\sum_{j=1}^{n} \mathbb{COV}[Y_{aj}, Y_{bj}]$$

Since our variables are also identically-distributed, we have...

$$= \frac{1}{n^2} n\mathbb{COV}[Y_a, Y_b]$$
$$= \frac{1}{n}\mathbb{COV}[Y_a, Y_b]$$

For example, here is Stata code.

```
use gss2018, clear
drop if class != 2 | missing(maeduc) | missing(paeduc)
gen malesspa = maeduc - paeduc // make the difference variable
sum malesspa
local n = r(N) // store the sample size for later
corr maeduc paeduc, cov // get the variance matrix for the variables
local cov = r(C)[2,1] // store the covariance in a local
local samplingvar = -2*`cov'*(1/`n')
    // turn it into the sampling covariance, or rather 2 times it
foreach var of varlist maeduc paeduc {
    qui sum `var'
    local SV = r(sd)^2/r(N)
    di `SV'
    local samplingvar = `samplingvar' + `SV'
    // get each sampling variance and put it into the formula
}
di `samplingvar'
local SE_paired = sqrt(`samplingvar') // get it into SE by taking root
di `SE_paired'
ttest maeduc == paeduc // now run Stata's paired t-test
local se_diff = r(se) // store its estimate of the SE of difference
ttest malesspa == 0 // finally just run a regular t-test on the difference variable
local se_diff_var = r(se) // store that SE

* Now let's tell Stata to check if these are different up to
* 10 decimal places (with computers, there is almost always
* tiny rounding error even with things that are almost
* exactly equal). -assert- produces no output if the claim is true

assert round(`SE_paired', 10) == round(`se_diff', 10)
assert round(`se_diff', 10) == round(`se_diff_var', 10)

* nice.
```